DISTANT SPEECH RECOGNITION IN REVERBERANT NOISY CONDITIONS EMPLOYING A MICROPHONE ARRAY

Juan A. Morales-Cordovilla, Martin Hagmüller, Hannes Pessentheiner, Gernot Kubin

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
{moralescordovilla,hagmueller,hannes.pessentheiner,gernot.kubin}@tugraz.at

ABSTRACT

This paper addresses the problem of distant speech recognition in reverberant noisy conditions employing a microphone array. We present a prototype system that can segment the utterances in real-time and generate robust ASR results off-line. The segmentation is carried out by a voice activity detector based on deep belief networks, the speaker localization by a position-pitch plane, and the enhancement by a novel combination of convex optimized beamforming and vector Taylor series compensation. All of the components are compared with other similar ones and justified in terms of word accuracy on a proposed database which simulates distant speech recognition in a home environment.

Index Terms— distant speech recognition; deep belief network voice activity detection; PoPi speaker localization; convex-optimized beamforming; vector Taylor series compensation; reverberant and noisy environment; natural mixing; German database.

1. INTRODUCTION

The distant interaction of a speaker with a dialogue system, which controls a home automation system, is a difficult challenge because of many reasons: the wake-up or attention of the system (distinction between human-human conversations and human-system commands), the change of the user accent in automatic speech recognition (ASR), and the degradation of the speech signal due to background noise or reverberation. Different challenges such as the recent REVERB [1] and projects such as CHIL, CHiME [2] and the current Distant-speech Interaction for Robust Home Applications (DIRHA, http://dirha.fbk.eu) have been introduced to solve this challenge.

To address the degradation problem, we propose the framework depicted in Fig. 1 which consists of state-of-the-art components. First, we acquire the sound signal with a 6-element star-shaped microphone array and segment it in multichannel utterances by means of a voice activity detector block based on deep belief networks (VAD-DBN). It is important to note that this block is the only one which processes the signal in real-time (i.e., the VAD output is provided, at least, as fast as the input samples) and that all the following blocks work off-line. This is the reason of situating at first place the VAD (the following blocks have a high computational cost and they can only process short segments of the whole temporal signal). Second, speaker localization (SLoc) based on the position and pitch (PoPi) plane estimates the spatial position of the multichannel utterance. Third, a convex-optimized beamformer (CVX-BF) provides a monaural enhanced signal which is compensated by vector Taylor series (VTS). Finally, we send the enhanced segment to an the ASR system.

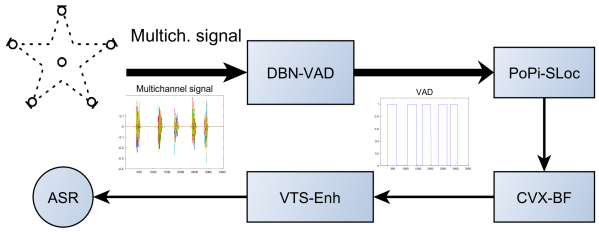

Fig. 1. Block diagram of the proposed system for distant speech recognition which consists of a 6-element microphone array, a voice activity detector based on deep belief networks (DBN-VAD), a speaker localizer based on the position and pitch plane (PoPi-SLoc), a convex-optimized beamformer (CVX-BF), a vector Taylor series enhancement (VTS-Enh) and an automatic speech recognizer (ASR).

In the literature we can find some proposals of similar systems [3, 4] but their applications are for games and meeting transcriptions respectively. This framework is an improved version of both, [5] and [6], because we do not use any true information such as the utterance segmentation and the speaker localization. The novelty of the paper is the analysis of the components to produce a positive synergy or cooperation. In order to do so, this paper also introduces a multichannel speech database which contains embedded clean signals of German commands for home automation control, contaminated with real room impulse responses and mixed in a 'natural' way [2] with real noise.

The paper is structured as follows: Section 2 presents the database and the ASR configuration. Section 3 justifies, mainly in terms of word accuracy (WAcc), the selection of the different components of the proposed system. Section 4 analyses in details the results and finally, section 5 summarizes the most important ideas presented in this paper together with future work.

2. EXPERIMENTAL FRAMEWORK

2.1. Embedded-DIRHA German Database

2.1.1. Description of the simulated environment

The proposed database, which we will call Embedded-DIRHA, refers to the 6-element star-shaped microphone array (1 center microphone and 5 on a pentagon on the same plane, with distance of 0.3 m to the center), placed on the ceiling of the living room of the ITEA apartment used by Fondazione Bruno Kessler (FBK) for the DIRHA

This work has been supported by the European project DIRHA FP7-ICT-2011-7-288121 and the Austrian Marshall Plan Foundation.

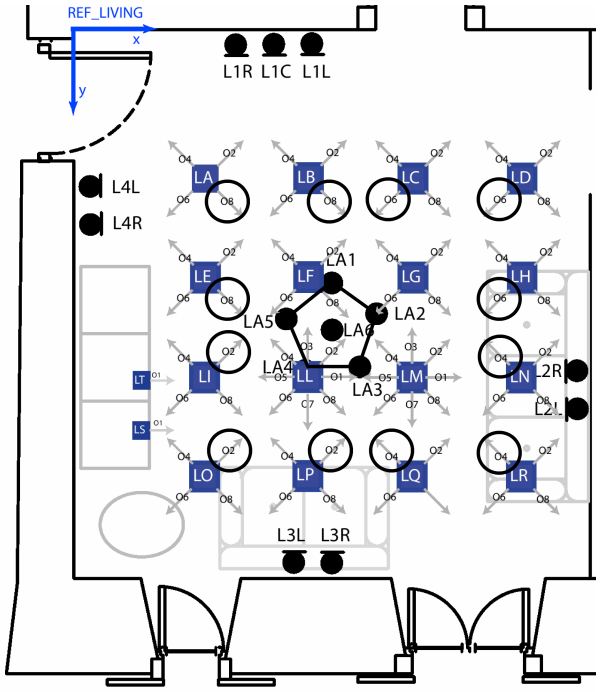

Fig. 2. Living room of the ITEA apartment of Fondazione Bruno Kessler (FBK) with the microphone array at the center and the 12 speaker positions employed in this work [provided by FBK].

project [7] (see Fig. 2). The selection of the geometry of this array is based on the trade-off between obtaining a uniform directivity and localization accuracy in any room position and having a small number of microphones. For the controllability of the experiments, the 12 speaker position-directions circled in Fig. 2 oriented to the center microphone are used. This provides us 72 (12 positions* 6 microphones) different impulse responses. The 2-D size of the room is 4.5x4.8 m and the average distance of these positions (1.5 m over the floor) to the center microphone is 2.4 m (enough to work with the wave plane assumption [8], and to consider distant speech recognition). The averaged reverberation time of the room is t_{60}=0.8 sec which is even higher than in REVERB challenge [1]. We have also 3 h of house noises (TV, washing-machine, children,...) recorded by this microphone array [7].

2.1.2. Test set

The database has a sampling frequency of 16 kHz and is divided into a test, training, and development set. The test set is furthermore divided in other three SNR sets: Clean, 10dB and 0dB. The last SNR (0dB) is very low for home environments ([1]) but we consider it to delimit better the performance of the proposed system. The Clean set consists of 57 6-channel-signals (from the star-shaped microphone array) of 45 sec duration on average. We call these signal clean embedded signals because they have concatenated or embedded 7 isolated speech utterances on average, separated randomly 0.5 to 5 sec and produced at different random positions by the same speaker. We use the impulse responses of Sec. 2.1.1 and 19 different speakers (10 male and 9 female) which gives 3 clean embedded signals per speaker. The 10dB and 0dB sets are obtained by natural mixing (Sec. 6.1) of the clean embedded signals with the 3 h of noise. We call this mix the embedded noisy signal. The isolated speech utterances are German read and spontaneous commands, and keywords of different lengths: 'Open the door', 'Turn on the bathroom fan and set it for thirty minutes', 'System!', .. extracted from the GRASS corpus [9]. In one SNR set we find 380 different isolated utterances and 296 different words.

2.1.3. Training and development sets

The clean training set contains 610 clean embedded signals (5046 isolated utterances) corresponding to 55 different-gender speakers: the above mentioned 19 GRASS [9] speakers (with different commands, keywords, and read sentences than test set) and 36 PHONDAT-1 [10] speakers. We mix two databases to do our recognition more robust to the speaker variation. To reduce the reverberation mismatch with the test set we employ the same 12 speaker positions. We also derive a multicondition training set by means of embedded noisy signals contaminated randomly at SNRs (Clean, 10dB and 0dB). The development set has 228 embedded signals corresponding to the 19 GRASS speakers of the multicondition set.

2.2. ASR system

The parameters of both, the front-end and the back-end, have been derived from the HTK-based recognizer of [11]. The front-end takes the enhanced signal and obtains mel frequency cepstrum coefficients (MFCCs) using: 16 kHz sampling frequency, frame shift and length of 10 and 32 ms, 1024 frequency bins, 26 mel channels and 13 cepstral coefficients with cepstral mean normalization. Delta and delta-delta features are also appended, obtaining a final feature vector with 39 components. The back-end is appropriate for our medium vocabulary size (296 words, Sec. 2.1.2). It employs a transcription of the training corpus based on 34 monophones (clustered from a previous 44 SAMPA-monophone transcription) to train triphone-HMMs. Each triphone is modeled by a HMM of 3 emitting states and 8 Gaussians/state. Although in the future we plan that the dialogue system selects the language model, now to avoid the difficulty of creating a deterministic grammar (due to the utterance diversity, Sec. 2.1.2) and because here we are more focused on the previous blocks to the ASR, we use a bigram trained on the test transcriptions. Note that the HMMs are trained with the center microphone signal of the training set without any enhancement (beamforming and VTS).

3. SYSTEM BLOCK DIAGRAM

3.1. Deep belief network voice activity detector

We propose the use of a voiced activity detector based on a deep belief network (DBN-VAD) because [12] shows that it performs better than many other VADs. Furthermore, due to its low computational cost in the feature extraction and decision stage, we can implement it in real-time. The implementation used here has a real-time of x1.08 [13]. The DBN-VAD takes the center microphone signal and provides the speech/non-speech decision for each frame from a feature vector. To train it we use the development set of Sec. 2.1.3 that provides us with around 1.400.000 feature vectors with their corresponding labels (derived from an energy-based VAD applied to the Clean set). We employ the same parameters as in Table II of [12] with a momentum of 0.9.

Unlike [12] each of our feature vectors has 282 coefficients (1-pitch; 26-Log-Mel-Spectrum and 13-MFCC with smooth time windows of 1, 9 and 17 frames centered around the frame under analysis;

12-LPC; 17-RASTA; and 135-AMS (Amplitude Modulation Spectrum)). These coefficients are completely derived from the pitch (from a simple extractor based on the maximum of the autocorrelation, a noise energy threshold equal to the amplitude mean, and a median smoothing to remove octave errors) and from the Log-Mel-Spectrum (Sec. 2.2). The final VAD decision is passed through a dilation morphological filter of 141 frames of total length (i.e. we grow in 70 frames the speech detection at the edges) to obtain the segmented utterance.

The averaged frame classification accuracy over the Clean, 10dB and 0dB test sets for different VADs is: True-VAD 100.00, DBN-VAD 84.39, the above mentioned energy-based-VAD 80.95, Extended-AFE-VAD [14] 80.30 and Extended-FE-VAD [15] 63.53 %. The corresponding WAcc (Sec. 6.2 to understand how to evaluate a VAD in terms of WAcc) results without any further enhancement are: 58.98, 57.39, 56.57, 55.24 and 47.03 %. These results clearly justify the use of the DBN-VAD.

3.2. Position-pitch speaker localizer

At each frame, we obtain a Position-Pitch (PoPi) plane [16] by means of the cross-correlation (cc) between the signal of the center microphone and the signal of one of the pentagon microphones. The PoPi value at one (pitch,angle) point is obtained by averaging the cc values which delay correspond to this point. The final PoPi plane for one frame is the average of the five PoPi planes of the pentagon. If we do it for all of the frames of the segmented utterance, we derive a PoPi function $PoPi(\alpha, f_0, t)$ that indicates the energy at a determined angle position α, pitch frequency f_0 and time frame t. In our case $\alpha = 0, 1, ..359^\circ$, $f_0 = 80, 81, ...220Hz$ and $t = 1, 2, ..fn$ (fn: frame number of the segmented utterance). If we consider that the speaker does not change his position during an utterance and for the sake of the simplicity, that the noise has a non-defined pitch along the utterance, we can estimate its position as the maximum of the marginalized PoPi function across the pitch and the time:

$$\hat{\alpha}_{speaker} = \arg\max_{\alpha}[\sum_{t=1}^{nf}\sum_{f_0=80}^{220} PoPi(\alpha, f_0, t)] \qquad (1)$$

where $\hat{\alpha}_{speaker}$ is the estimated azimuth angle of the speaker relative to the center microphone. The elevation is not estimated because we assume always 1.5 m over the floor (Sec. 2.1.1). A similar marginalized approach is used to localize the speaker by means of the SrpPhat function [17]. The averaged accuracy of the speaker angle estimation (Sec. 6.3) over the test set for different speaker localizers when we use the true utterance segmentation is: True-SLoc 100, PoPi-SLoc 95.82 and SrpPhat-SLoc 93.61 %. The angle accuracy when we use the DBN-VAD segmentation is respectively: 99.93, 95.20 and 92.97 %. The corresponding WAcc when we use these positions to enhance the signal with the CVX-BF (Sec. 3.3) is: 66.11, 65.37 and 65.65 for the true segmentation; 64.04, 63.54 and 63.79 % for the DBN-VAD segmentation. Although the SrpPhat gives a bit better result in WAcc than the PoPi-SLoc, we choose the latter because it has more accuracy, has a bit lower computational cost and fits better in our future work (Sec. 5).

3.3. Convex optimized beamformer

The main difference between the convex optimized (CVX) [6, 5] and the delay and sum (DS) [8] beamforming is that the former one can better approximates the desired beampattern along all the different frequencies and spatial directions. This allows the 3D optimization of the main lobe and also reduces the influence of ceiling and floor reflections. To avoid computing new beamforming coefficients for every estimated position, we only compute it for the 12 speaker positions and we use the coefficients corresponding to the closest position. The WAcc when we use the true utterance segmentation and true speaker localization for different beamformers is: CVX-BF 66.11, DS-BF 64.69 and No-BF 58.98 %. The corresponding WAcc when we use DBN-VAD segmention and PoPi-SLoc is: 63.54, 61.93 and 57.39 %. All of these results justify the use of the CVX-BF.

3.4. Vector Taylor series enhancement and noise estimation

The reason for using VTS rather than other methods, such as marginalization of missing data [18], is that the representation of the clean estimated signal can be in the cepstral domain, which is a more appropriate representation for medium vocabulary recognition (Sec. 2.2). Given the noisy feature vector $\mathbf{y}_t$ at frame time t, a 0th order VTS estimates the clean feature vector $\hat{\mathbf{x}}_t$ as follows [19]:

$$\hat{\mathbf{x}}_t = \mathbf{y}_t - \sum_{k=1}^{K} P(k|\mathbf{y}_t)\boldsymbol{g}\left(\mu_X^{(k)}, \hat{\mathbf{n}}_t\right), \qquad (2)$$

where $\hat{\mathbf{n}}_t$ is the noise estimation and $\boldsymbol{g}(\mathbf{x}, \mathbf{n}) = \log(\mathbf{1}+\exp(\mathbf{n}-\mathbf{x}))$. $P(k|\mathbf{y}_t)$ and $\mu_X^{(k)}$ are the component probability and mean of a Gaussian mixture model (GMM) of clean speech with K components. Mention that we use $K = 256$ GMM components, covariance matrix of the noise estimation and 26-Log-Mel-Spectrum feature vector (Sec. 2.2). We use a noise estimate based on the first-last-frames (FLFr-Noise, [6]). This estimation assumes that the first and last 20 frames of the segmented signal correspond to noise and these frames are used to estimate the Log-Mel-Spectrum noise (and its corresponding covariance matrix) by means of a linear interpolation over the remaining frames. The WAcc when we use the true utterance segmentation, true speaker localizer and CVX-BF for different noise estimations on VTS is: True-Noise 78.22, FLFr-Noise 70.96, Min-Statistics-Noise [20] 68.76 , MMSE-noise [21] 51.92 %. The corresponding WAcc with DBN-VAD segmentation, PoPi-SLoc and CVX-BF is: 71.57, 65.69, 63.80 and 48.23 %. In addition, it is worth to mention that we also tried other kinds of compensation algorithms such as binary imputation [6]. This imputation provides worse results than VTS and needs tuning of the threshold of the missing data mask. All of these results justify the use of VTS-Enh with FLFr-Noise.

4. ANALYSIS OF THE RESULTS

Table 1 summarizes the most important WAcc results presented in the paper. If we pay attention to the average column, we can see that the two proposed enhancements produce a positive synergy: we start with 57.39, then the addition of the CVX-BF gives 63.54 and the addition of VTS 65.69 %. This is our best result without using any true information. This synergy can be even bigger with a more precise noise estimation (see the 74.57 that we can reach with the true noise on VTS). We can also reach 73.05, without using any true information, if we apply the enhancements on the multicondition training set of Sec. 2.1.3. Without the enhancements, we reach 66.40. These two results are better than their respective 65.69 and 57.39 because the mismatch between the trained HMMs and the test sentences is reduced. To measure the robustness of the system to speaker position changes we create a new test set, as in Sec. 2.1.2, but using the positions not included in the training, i.e., the closest to the array: LF/08, LG06, LM/04 and LL/02 (Fig. 2). Its averaged result

Table 1. Word accuracies (WAcc, %) obtained by different configurations of the proposed system tested over the presented Embedded-DIRHA database for the different SNR sets.

Utterance segmentation	Speaker localiz. and beamforming	Type of noise and enhacement	Clean	10 dB	0 dB	Average
True-VAD	No	No	81.33	61.31	34.29	58.98
DBN-VAD	No	No	81.49	61.82	28.87	57.39
True-VAD	True-SLoc and CVX	No	83.54	68.95	45.83	66.11
True-VAD	PoPi-SLoc and CVX	No	83.47	68.55	44.08	65.37
DBN-VAD	True-SLoc and CVX	No	84.81	66.40	40.90	64.04
DBN-VAD	PoPi-SLoc and CVX	No	84.73	65.92	39.98	63.54
True-VAD	True-SLoc and CVX	True-N and VTS	84.81	80.22	69.62	78.22
True-VAD	PoPi-SLoc and CVX	FLFr-N and VTS	84.81	74.68	53.40	70.96
DBN-VAD	True-SLoc and CVX	True-N and VTS	83.28	78.29	61.15	74.57
DBN-VAD	PoPi-SLoc and CVX	FLFr-N and VTS	84.10	69.86	43.12	65.69

without retraining the GMMs of VTS and HMMs to the new possible reverberations is 61.65. This result is not very different from the 65.69 of the farther positions. This can be explained by the fact that, although on the 2-D plane the distance to the array seems to change a lot, in 3-D it is really almost the same distance and consequently the same reverberation time. Finally, if we observe the Clean column we see that, despite having a medium size lexicon (Sec. 2.1.2), the performance is low (around 84 for our best system) compared with the one (95) obtained with a similar database such as [6]. This is because of the errors in the keywords for which the bigram does not help for their predictions (the WAcc of the commands is 87 and of the key words 78).

5. CONCLUSION AND FUTURE WORK

This paper presented a system for distant speech recognition in reverberant and noisy conditions, intended to control a room with commands by means of signal recorded by a star-shaped microphone array. The proposed system has been an improved version of the systems presented in [6, 5]. This improvement has consisted of the creation of a prototype which can segment utterances in real-time, localize the speaker and generate robust ASR results off-line by means of a novel combination of CVX-BF with VTS enhancements. The selection of the different components has been carefully justified in terms of WAcc, and another metrics such as VAD classification and angle accuracy. To do so, we have proposed a database that simulates distant speech recognition in a home environment. We have improved the results even more by means of multicondition training and show that the system is robust to speaker position changes. As future work we plan to explore these two ideas: 1) the fusion of some of the components, such as the PoPi-localizer with the VAD by means of the pitch [22] and 2) the extension of the system to all the rooms of the ITEA apartment by using the network of microphone arrays.

6. APPENDIX

6.1. Natural mixing

We compute the target SNR as:

$$SNR = 10\log_{10}\frac{Ex_{central}}{En_{central}}(dB) \quad (3)$$

where $Ex_{central}$ and $En_{central}$ are the whole energy of the central microphone of the clean embedded signal and of a noise segment with same length respectively. Natural mixing [2] means that we randomly select the noise segment that provides our target SNR (within an error of $\pm$ 1.5 dB) without modifying the gain of both, the clean and the noise signals. If no noise segment which fits the target SNR is found, all channels of the clean signal are multiplied by a gain (which depends on the closest found SNR to the target SNR) to find at least an appropriate noise segment. In addition, if the final mix shows saturation in some of the channels we multiply all of them by a common factor to avoid this problem.

6.2. Transcription association

In order to evaluate the VAD in terms of WAcc, we associate the word transcription to the estimated segments as follows: If we have a larger number of estimated than of true segments, the associated transcription of the estimated segment starts with the transcription of the nearest true segment and finishes when we cover all the true transcriptions. The estimated segments that remain without transcription are not recognized. If we have a lower number of estimated than of true segments, we do the same, but the true segments, which remain without associated estimated segments, are recognized as white noise with the true transcription. This association guarantees that, independently of the used VAD, we always recognize the same number of segments.

6.3. Angle accuracy

We compute the accuracy of the speaker estimated angle as:

$$Acc_{angle} = 100(1 - d_{angle}(\hat{\alpha}, \alpha)/180) \quad (4)$$

where $\hat{\alpha}$ and α are the estimated and true angles in degrees and d_{angle} is the minimum angle difference between two angles (note that its value is always in the interval $[0, 180^{\circ}[$).

REFERENCES

[1] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakataniand, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE WASPAA*, 2013.

[2] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: A resource and a challenge for computational hearing in multisource environments," in *Interspeech*, 2010.

[3] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE, Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, 2012.

[4] C. Fredouille and N. Evans, "The lia rt07 speaker diarization system," *Lecture Notes in Computer Science (MTPH). Springer*, vol. 4625, pp. 520–532, 2008.

[5] H. Pessentheiner, S. Petrik, and H. Romsdorfer, "Beamforming using uniform circular arrays for distant speech recognition in reverberant environments and double-talk scenarios," in *Interspeech*, 2012.

[6] J. A. Morales-Cordovilla, H. Pessentheiner, M. Hagmüller, P. Mowlaee, F. Pernkopf, and G. Kubin, "A German distant speech recognizer based on 3D beamforming and harmonic missing data mask," in *AIA-DAGA*, 2013.

[7] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmüller, and P. Maragos, "The DIRHA simulated corpus," in *LREC*, 2014.

[8] I. Tashev, *Sound Capture and Processing: Practical Approaches*, John Wiley and Sons, 2009.

[9] B. Schuppler, M. Hagmüller, J. A. Morales-Cordovilla, and H. Pessentheiner, "GRASS: The Graz corpus of read and spontaneous speech," in *LREC*, 2014.

[10] F. Schiel and A. Baumann, "Phondat 1, corpus v.3.4.," Tech. Rep., Bavarian Archive for Speech Signals (BAS), 2006.

[11] H. G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends of large vocabulary task," Tech. Rep., ETSI STQ-Aurora DSR, 2002.

[12] X. L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. ASLP*, vol. 21, no. 4, pp. 697–710, 2013.

[13] J. A. Morales-Cordovilla, H. Pessentheiner, M. Hagmüller, and G. Kubin, "Room localization for distant speech recognition," in *Interspeech*, 2014.

[14] ES 202 212 v1.1.2, *Extended advanced front-end feature extraction algorithm*, ETSI, November 2005.

[15] ES 202 211 v1.1.1, *Extended front-end feature extraction algorithm*, ETSI, July 2001.

[16] T. Habib and H. Romsdorfer, "Auditory inspired methods for localization of multiple concurrent speakers," *Computer Speech & Language*, vol. 27, no. 3, pp. 634–659, 2012.

[17] J. H. Dibiase, *A high accuracy, low latency technique for talker localization in reverberant environments using microphone arrays*, Ph.D. thesis, Brown University, 2000.

[18] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

[19] J. C. Segura, A. de la Torre, M. C. Benitez, and A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks," in *Eurospeech*, 2001.

[20] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. SAP*, vol. 9, no. 5, pp. 504–512, 2001.

[21] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans ASLP*, vol. 20, no. 4, pp. 1383–1393, 2012.

[22] N. Ma, J. Barker, H. Christensen, and P. Green, "Binaural cues for fragment-based speech recognition in reverberant multi-source environments.," in *Interspeech*, 2011.